Data Science Bootcamp Capstone

By Chan Shao Mun

Problem Statement

- Credit card default refers to the scenario where the owner of the credit card fails to make the minimum payment on a credit card by a due date.
- The limited ability of commercial banks to identify high default risk credit clients at the point of credit card application is a persistent issue that contributes to the lender's financial losses from credit risk.

PROJECT AIM

This project aims to test if various customer information factors like gender, car ownership, property ownership, annual income, education status, number of family members/children and more could be used to predict the likelihood of a client defaulting on their credit card loans.

Pre-Analysis : Data Cleaning and Transformation

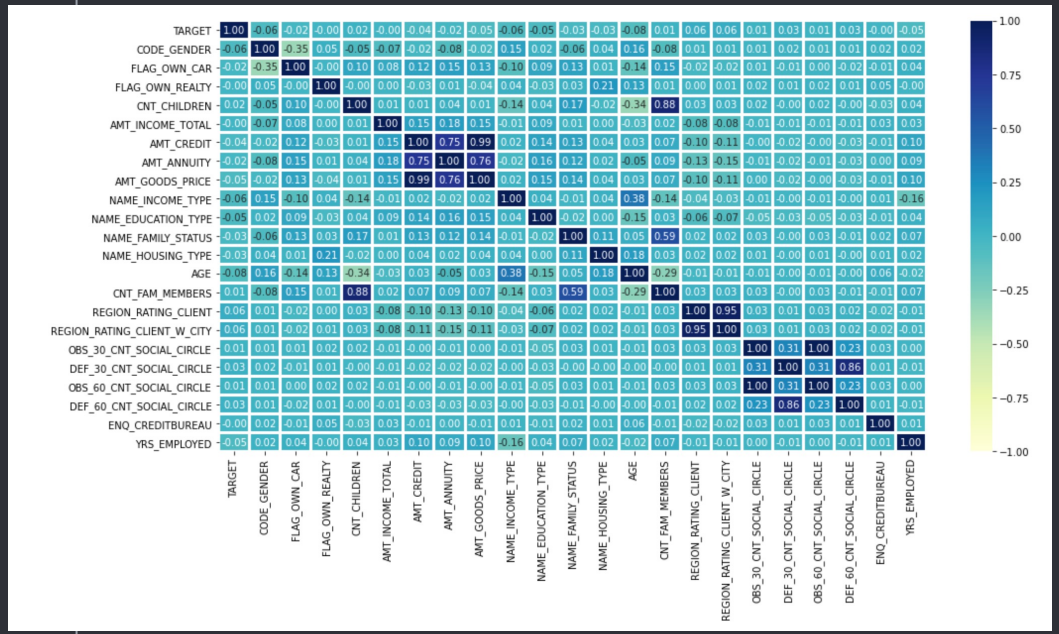
- Converted client's age in days to age in years
- There were 6 columns representing how many enquiries were made to the Credit Bureau within different time frames (1 hour, 1 day, 1 week) prior the application – summed all columns into 1 column representing number of enquiries made within the period of 1 year before the application
- Dataset represented number of days that the client has been employed as a negative number; Clients who were pensioners or unemployed were given the positive value of 365243 – Changed their values to zero instead to better represent their length of employment.
- Checked and removed duplicates
- Removed entries with null values ("XNA", "Unknown")
- Narrowed down the dataset from 122 columns to 23 columns by removing columns that had too many categorical values or were irrelevant to analysis.

Selected variables in dataset

0	TARGET	278217 non-null	int64
1	CODE_GENDER	278217 non-null	object
2	FLAG_OWN_CAR	278217 non-null	object
3	FLAG_OWN_REALTY	278217 non-null	object
4	CNT_CHILDREN	278217 non-null	int64
5	AMT_INCOME_TOTAL	278217 non-null	float64
6	AMT_CREDIT	278217 non-null	float64
7	AMT_ANNUITY	278217 non-null	float64
8	AMT_GOODS_PRICE	278217 non-null	float64
9	NAME_INCOME_TYPE	278217 non-null	object
10	NAME_EDUCATION_TYPE	278217 non-null	object
11	NAME_FAMILY_STATUS	278217 non-null	object
12	NAME_HOUSING_TYPE	278217 non-null	object
13	AGE	278217 non-null	float64
14	CNT_FAM_MEMBERS	278217 non-null	float64
15	REGION_RATING_CLIENT	278217 non-null	int64
16	REGION_RATING_CLIENT_W_CITY	278217 non-null	int64
17	OBS_30_CNT_SOCIAL_CIRCLE	278217 non-null	float64
18	DEF_30_CNT_SOCIAL_CIRCLE	278217 non-null	float64
19	OBS_60_CNT_SOCIAL_CIRCLE	278217 non-null	float64
20	DEF_60_CNT_SOCIAL_CIRCLE	278217 non-null	float64
21	ENQ_CREDITBUREAU	278217 non-null	float64
22	YRS_EMPLOYED	278217 non-null	float64

- The "TARGET" column is the variable that we would like to predict.
  - 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample
  - 0 - all other cases

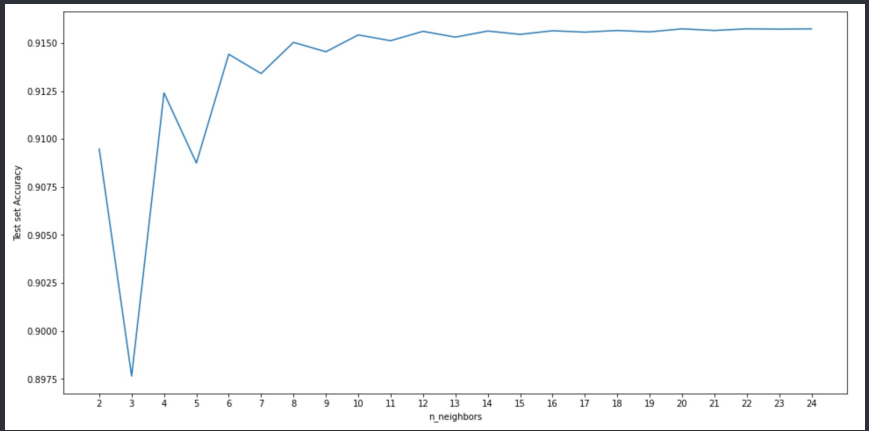
Correlation Matrix



From the correlation matrix above, we can see that there are no strong correlations between TARGET and any of the other features.

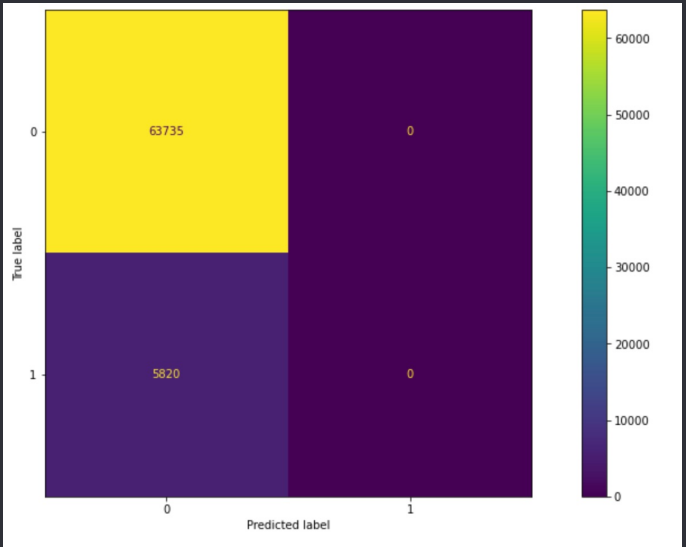
We can observe some strong positive strong correlations between the other features but none of them are a surprise as the variables tend to be derived/subsets of each other. *I.e. children will be counted under client's family members, loan annuity amount will be derived from credit amount of the loan, credit amount of loan given is likely to be dependent on the price of goods for which the loan is given.*

KNN



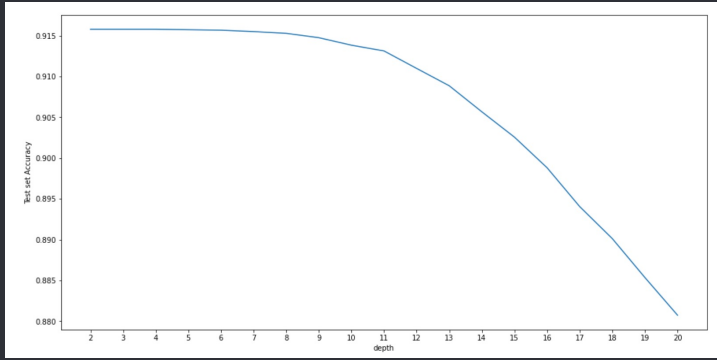
- Iterated through n_neighbors from 2 to 24.
- The accuracy of the kNN model starts to plateau around n_neighbours = 12.
- Accuracy score at n_neighbors = 24: 0.9157357486880886

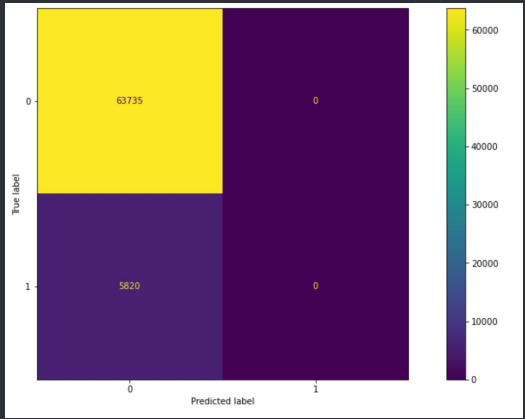
Logistic Regression



- Ran logistic regression model at default parameters.
- Accuracy score: 0.9157932571346417
- Confusion Matrix: 63735 True negatives, 5820 False negatives, 0 true positives, 0 false positives

Decision Tree





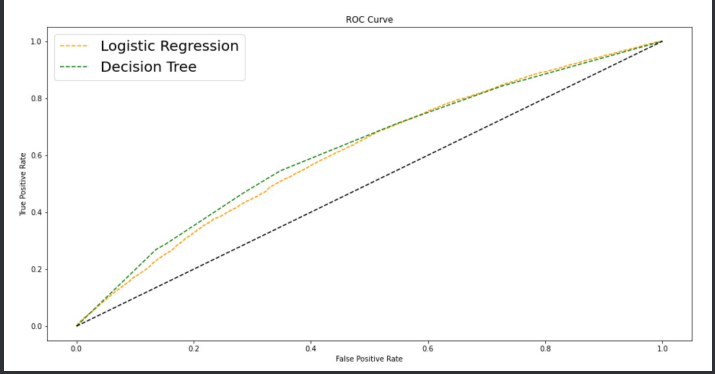
- Iterated through depth = 2 to 20
- The highest accuracy score of 0.9157932571346417 occurs at depth = 2 to 4 and then proceeds to decrease due to overfitting of the model
- Accuracy score: 0.9157932571346417
- Confusion Matrix: 63735 True negatives, 5820 False negatives, 0 true positives, 0 false positives

Model Performance

Model	Accuracy Score
KNN	0.9157357486880886
Logistic Regression	0.9157932571346417
Decision Tree	0.9157932571346417

- Logistic Regression and Decision Tree models performed slightly better than the KNN model.

Comparison of Logistic Regression and Decision Tree



Model	AUC Score
Logistic Regression	0.6115595718634153
Decision Tree	0.6214248956091548

- Ran the two models again on a different train-test split.
- Both produced the same accuracy score again

Model	Accuracy Score
Logistic Regression	0.9163252102652577
Decision Tree	0.9163252102652577

- Calculated the AUC scores for both models. Decision Tree performed marginally better

Conclusions

- The logistic regression model and decision tree model performed better than the KNN model.
- Both the logistic regression model and decision tree model produced the same accuracy score, even when tested on different test sets.
- When comparing the AUC scores of both models, the decision tree model (0.621) performed marginally better than the logistic regression model (0.612).
- However, after observing the confusion matrices, I would say that both models are not good at predicting high default risk clients at all as both of them did not predict any positives (TARGET ==1) and classified all clients as negative (TARGET == 0; no-risk). Much improvement on the models are hence needed.

Limitations and Recommendations

- Limitation: I narrowed down the analysis from the 122 columns available in the dataset to 23 columns - the accuracy of the classification models could possibly have been improved if I used more variables in the analysis dataset.

- A future recommendation could be to use Gridsearch to further finetune and optimize the parameters of the decision tree model.
- Another recommendation would be to re-evaluate if there are other factors that we did not include in the analysis (e.g. if client submitted all documents, default or approval rate on previous applications) that could help improve the accuracy of the model.

Source of Dataset

- https://www.kaggle.com/datasets/mishra5001/credit-card

Thank you!